High throughput sequencing of full-length SSU rRNA sequences from complex microbial communities without primer bias and how it affects our ability to study microbial ecology

Dueholm, Morten Simonsen; Karst, Søren Michael; McIlroy, Simon Jon; Kirkegaard, Rasmus Hansen; Nielsen, Per Halkjær; Albertsen, Mads

Publication date:
2017

High Throughput Sequencing of Full-length SSU rRNA Sequences from Complex Microbial Communities without Primer Bias and how it Affects our Ability to Study Microbial Ecology

M. S. Dueholm, S. M. Karst, S. J. McIlroy, R.H. Kirkegaard, P. H. Nielsen, and M. Albertsen

Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, Denmark

Introduction

The small subunit (SSU) ribosomal RNA (rRNA) genes have been used to study of microbial diversity and evolution for the last 30 years. Today, databases containing full-length SSU rRNA reference sequences remains fundamental for many core analyses in microbial ecology, such as community profiling using 16S/18S rRNA amplicon sequencing and in-situ studies based on fluorescence in situ hybridization (FISH) microscopy. The quality of the data produced relies heavily on the reference databases used, and it is widely recognized that the current databases are underpopulated, ecosystem skewed, and subject to primer bias. Here we present a method that combines reverse transcription of polyadenylated full-length SSU rRNA molecules with Illumina based synthetic long-read sequencing to obtain high quality, full-length SSU rRNA sequences in a high throughput manner. We applied the approach to complex samples from seven different ecosystems and obtained more than 1,000,000 SSU rRNA gene sequences from all domains of life with an estimated raw error rate of 0.17%. We observed a high fraction of novel diversity including several deeply branching phylum level lineages. Here we describe how the method works and demonstrate how the access to comprehensive ecosystem specific SSU databases affect our ability to study microbial ecology.

Method overview and evaluation

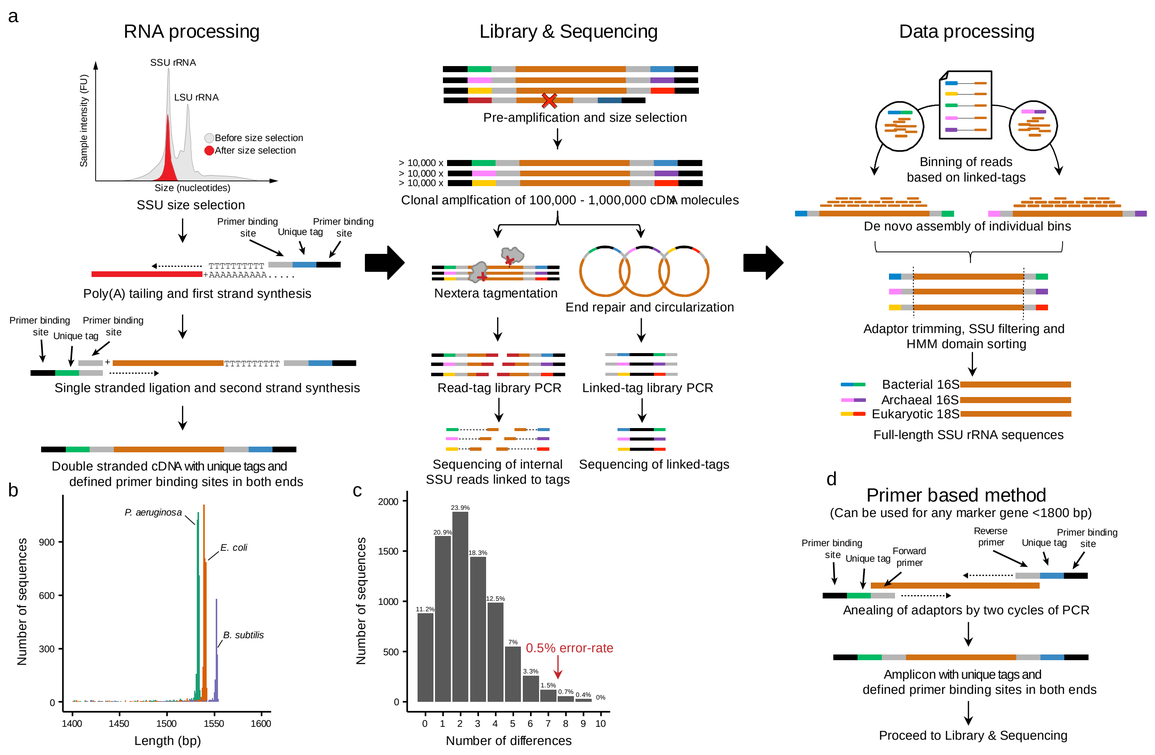

Figure 1. Overview and validation of the full-length SSU rRNA sequencing protocol. a, Schematic overview of the preparation of full-length SSU rRNA gene sequences from total community RNA. First, SSU rRNA is enriched from extracted total community RNA using size selection. The enriched SSU rRNA is polyadenylated and the polyA-tail used as a generic priming site for first strand synthesis (reverse transcription). Another generic priming site is added to the other end of the cDNA molecule by single stranded ligation and used for second strand synthesis. PCR adaptors used for first and second strand synthesis contain unique tags (green and blue), which, in combination, become the unique "linked-tags" of the molecules. The adaptors also contain generic priming sites that are used for PCR amplification of the tagged molecules. The amplicons are size selected to remove incomplete or truncated products. A defined number of full-length SSU rRNA amplicons are amplified with PCR to generate >10,000 copies of each uniquely tagged amplicon. The clonal amplicon library is split in two and used for preparing a read-tag library and a linked-tag library. The read-tag library is prepared by fragmenting the full-length SSU rRNA amplicons using Illumina Nextera tagmentation and library preparation. The resulting sequencing outcome is an internal SSU rRNA fragment read connected to a single unique tag read. The linked-tag library is prepared by circularizing full-length SSU rRNA amplicons to physically link the tags in close proximity. PCR is used to amplify the linked-tags, which are then identified with sequencing. The linked-tags are used to bin all SSU rRNA fragment tag-reads originating from the same parent molecule. De novo assembly is used to recreate the parent SSU rRNA gene sequence. The resulting sequences are finally, trimmed for adaptors, filtered and classified by HMM domain soting. b, Size distribution of assembled SSU rRNA gene sequences from the mock community. c, Error count distribution for raw SSU rRNA sequences from the mock community (Numbers indicate percent of all 16S rRNA gene sequences). d, Adaptation of the method to be used with primers. Generic primer binding sites and unique tags are added to each end of the target by two round of PCR and the resulting tagged amplicons are treated in the same way as the tagged cDNA molecules.

Application to environmental samples

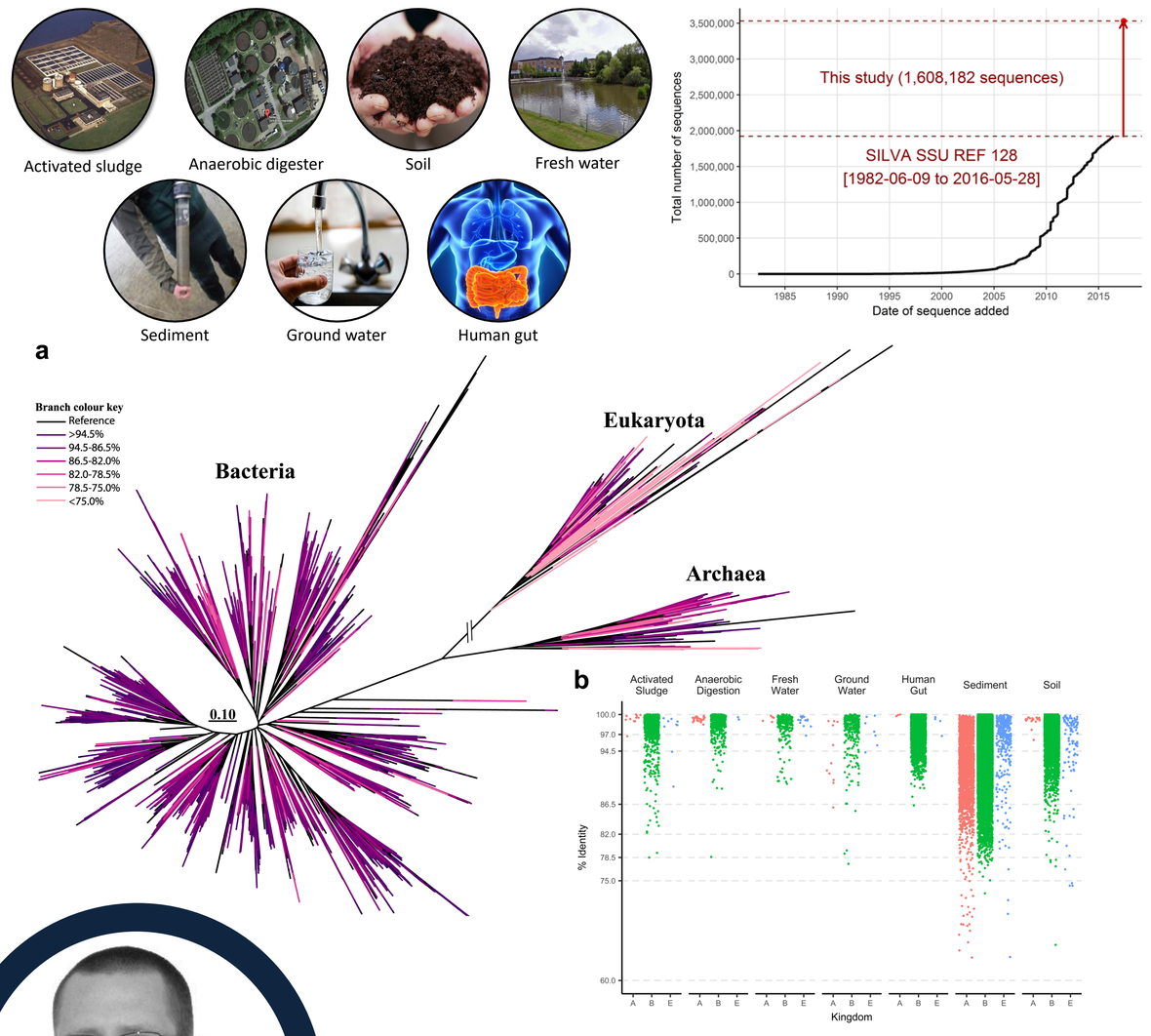

Figure 2. Coverage of the tree of life. a, Maximum-likelihood phylogenetic tree showing coverage of the tree of life. The tree includes all OTUs clustered at 97% generated in this study, their closest match in the Silva SSU NR99 v. 128 database and the reference set from the recent Tree of Life article (Hug et al., 2016). Hypervariable regions were masked with a 5% positional conservation filter, giving 1698 alignment positions, and the tree calculated using FastTree v. 2.1.3 SSE3 (Price et al., 2010). Reference sequences appear black whilst those generated in the current study are color coded based on their similarity to existing database sequences. b, The percent identity of SSU rRNA gene sequences in the samples compared to their closest relatives in the SILVA database. A: Archaea, B: Bacteria, E: Eukaryota.

Novel phylum-level archaeal lineages (DAS1-8)

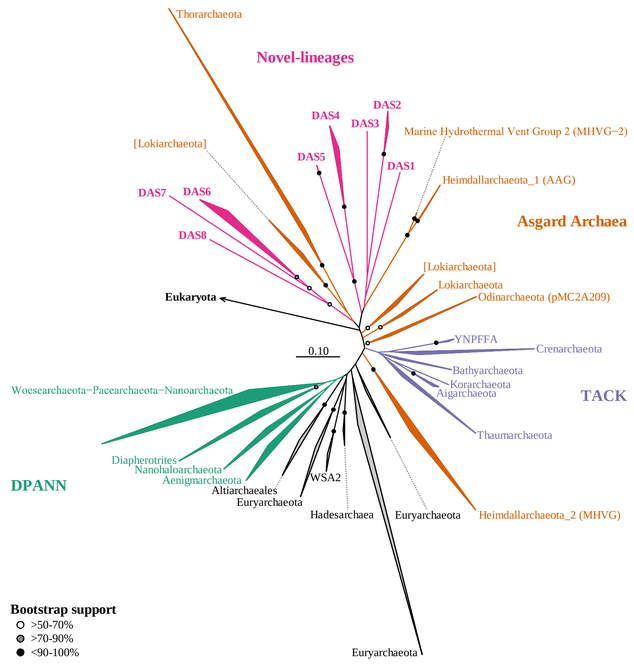

Figure 3. Discovery of novel archaeal diversity. a, Maximum-likelihood phylogenetic tree showing coverage of the domain Archaea. The tree contains archaeal OTUs clustered at 97% generated in this study that have ≤82% similarity to any database sequence, their closest match in the Silva SSU NR99 v. 128 database, the archaeal reference set from the recent 'Tree of Life' article (Hug et al., 2016) and representatives of the newly described Asgard Archaea super phylum (incl. Thorarchaeote SMTZ1-83, Odinarchaeote LCB_4 and, Heimdallarchaeote LC_3 and AB_125)(Zaremba-Niedzwiedzka et al., 2017). Hypervariable regions were masked with a 50% positional conservation filter, giving 1078 alignment positions, and the tree calculated using RAxML v. 8.2.10 (Stamatakis, 2014). Bootstrap support is shown where greater than 50% based on 'rapid bootstrap' analysis with 1000 iterations. Clade names and clustering are based on the position of reference sequences. Clades names enclosed in square brackets do not include a genome or pure culture reference sequence – being based on classification of reference sequences in the Silva v. 1.28 taxonomy. Clades are coloured by their affiliation to archaeal superphyla. Novel lineages detected in this study appear magenta.

The hidden diversity of the rare biosphere

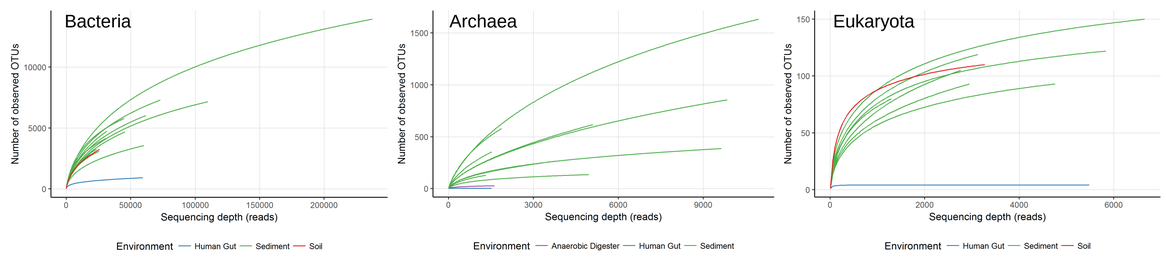

Figure 4. Rarefaction curves for the different samples split based on kingdom. Deep sequencing of sediment samples demonstrates an amazing diversity of the rare biosphere in theses samples with 1 out of 50 sequencing reads representing a novel OTU after 200,000 reads. The observed diversity is not a result of sequencing error or chimera, which is evident from the flattening of the eukaryotic data for the human gut samples.

Conclusions and perspectives

- Millions of primer-free, high-quality full-length SSU rRNA sequences from all domains of life
- A fully populated tree of life is [maybe] within reach
- Novel undescribed phylum-level archaeal linages discovered
- Primer-version compatible with any marker gene [<1800 bp]
- Partial 23S/28S LSU genes also obtained, doubling the potential for design of new primers and probes
- More robust evaluation and design of primers and probes
- Eco-system specific databases will enable highly specific primers and probes
- Enhance all *in situ* studies that rely on primers and probes e.g. FISH
- Guide efforts to obtain genomes from undescribed branches of the tree of life

Interactive protocols at protocols.io

Primer-free version

- dx.doi.org/10.17504/protocols.io.h2rb8d6

Primer-based version

- dx.doi.org/10.17504/protocols.io.h2sb8ee

Morten Simonsen Dueholm md@bio.aau.dk Center for Microbial Communities, Aalborg University